Review

Juergen Kratzsch, Nikola A. Baumann, Ferruccio Ceriotti, Zhong X. Lu, Matthias Schott, Antonius E. van Herwaarden, José Gilberto Henriques Vieira, Dusanka Kasapic and Luca Giovanella*

Global FT4 immunoassay standardization: an expert opinion review

https://doi.org/10.1515/cclm-2020-1696
Received November 11, 2020; accepted December 13, 2020;
published online December 23, 2020

Abstract

Objectives: Results can vary between different free thyroxine (FT4) assays; global standardization would improve comparability of results between laboratories, allowing development of common clinical decision limits in evidence-based guidelines.

Content: We summarize the path to standardization of FT4 assays, and challenges associated with FT4 testing in special populations, including the need for collaborative efforts toward establishing population-specific reference intervals. The International Federation of Clinical Chemistry and Laboratory Medicine Committee for Standardization of Thyroid Function Tests has undertaken FT4 immunoassay method comparison and recalibration studies and developed a reference measurement procedure that is currently being validated. Further studies are needed to establish common reference intervals/clinical decision limits. Standardization of FT4 assays will change test results substantially; therefore, a major education program will be required to ensure stakeholders are aware of the benefits of FT4 standardization, planned transition procedure, and potential clinical impact of the changes. Assay recalibration by manufacturers and approval process simplification by regulatory authorities will help minimize the clinical impact of standardization.

Summary: Significant progress has been made toward standardization of FT4 testing, but technical and logistical challenges remain.

Outlook: Collaborative efforts by manufacturers, laboratories, and clinicians are required to achieve successful global standardization of the FT4 assays.

Keywords: free thyroxine; immunoassay; reference value; standardization; thyroid.

Introduction

Thyroid function tests are among the most frequently requested laboratory procedures [1, 2]; therefore, reliable assays are crucial for optimal patient care. While a normal thyroid stimulating hormone (TSH) concentration is usually adequate to exclude thyroid disease in asymptomatic patients [3], measurement of free thyroxine (FT4) is relevant to differentiate subclinical from overt hyperthyroidism [4, 5] or hypothyroidism [6, 7], and to investigate suspicious abnormal TSH secretion [2, 4, 8, 9], TSH-secreting pituitary tumor [10], or thyroid hormone resistance [11]. FT4 can be measured using various methods in the laboratory. As 99.98% of thyroxine is bound to proteins [12], FT4 assays must be able to accurately measure the 0.02% of biologically active thyroxine that exists as FT4.

***Corresponding author: Luca Giovanella**, PhD, Clinic for Nuclear Medicine and Competence Centre for Thyroid Diseases, Ente Ospedaliero Cantonale, Bellinzona, Switzerland; and University Hospital and University of Zurich, Via A. Gallino 12, 6500 Bellinzona, Switzerland, Phone: +41 91 811 86 72, Fax: +41 91 811 85 02, E-mail: luca.giovanella@eoc.ch. https://orcid.org/0000-0003-0230-0974
Juergen Kratzsch, Institute for Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital, University of Leipzig, Leipzig, Germany
Nikola A. Baumann, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
Ferruccio Ceriotti, Clinical Laboratory, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy. https://orcid.org/0000-0002-0958-5354
Zhong X. Lu, Department of Medicine, Monash University, Victoria, Australia
Matthias Schott, Division for Specific Endocrinology, Medical Faculty, University of Düsseldorf, Düsseldorf, Germany
Antonius E. van Herwaarden, Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen, The Netherlands
José Gilberto Henriques Vieira, Division of Endocrinology, EPM/UNIFESP and Grupo Fleury, São Paulo, Brazil
Dusanka Kasapic, Roche Diagnostics International Ltd., Rotkreuz, Switzerland

FT4 measurement methods are generally classified as direct or indirect. Direct methods employ physical separation of the free hormone from the protein-bound hormone in the sample using techniques such as equilibrium dialysis (ED) or filtration. The separated FT4 fraction is then measured, usually by liquid chromatography (LC) with tandem mass spectrometry (MS) [13]. Indirect methods are widely used in clinical laboratories and measure FT4 on automated immunoassay platforms, which utilize different competitive assay formats to selectively measure non-bound FT4 without disrupting the protein-bound thyroxine [14].

FT4 immunoassays are known to present challenges associated with measuring low concentrations of biologically active free thyroid hormone relative to total hormone concentration. The analytical challenge is to measure FT4 itself without disturbing the equilibrium of thyroxine and its binding proteins. With current FT4 immunoassays, changes in binding protein concentrations can significantly influence the test results in a method-dependent manner, and potential inaccuracies at low or high concentrations commonly observed in individuals with hypo- and hyperthyroid may lead to misclassification of patients [15]. The inverse log-linear relationship between FT4 and TSH was significantly better when FT4 was measured directly by LC-MS/MS than by an indirect immunoassay, especially in subjects with normal TSH values [16]. These limitations are more profound in conditions where binding protein concentrations can be significantly altered, such as acute illness, pregnancy [17], and hereditary variants in the structure of thyroxin-binding globulin, albumin, or transthyretin [18].

In addition to these analytical limitations, FT4 test results are known to vary between assays from different manufacturers. Data from the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee for Standardization of Thyroid Function Tests (C-STFT) have demonstrated differences in results across 13 FT4 assays [19]. For example, the median (range) bias was −24% (−14% to −42%) for samples in the 9–27 pmol/L range (prior to recalibration to a reference measurement procedure [RMP]) [19]. This finding is also evidenced in a study of FT4 result distributions of more than 600,000 routine clinical samples by two different FT4 assays (Figure 1, unpublished data, courtesy of José Gilberto Henriques Vieira).

Global standardization would allow traceability of results back to a common standard and improve the comparability of FT4 measurements between assays, providing the basis for the development of common reference intervals, decision limits, and standards of medical care. However, there are substantial challenges in achieving global standardization. It requires establishment of an RMP, recalibration of existing assays by all manufacturers and going through comprehensive approval process by regulatory bodies, and determination of reference intervals. In addition, commutability of the FT4 reference material in different analytical platforms compared with clinical samples from specific patient groups with altered binding protein concentrations, such as pregnant women, would also need to be examined post-standardization. Standardization would significantly change FT4 numeric results in many laboratories; therefore, a substantial education program would be required to explain the changes to laboratories, clinicians, and patients [20]. This review was developed following an expert panel meeting of clinicians and laboratory specialists who convened to discuss the current landscape in FT4 testing, and the potential advantages and challenges associated with global FT4 standardization.

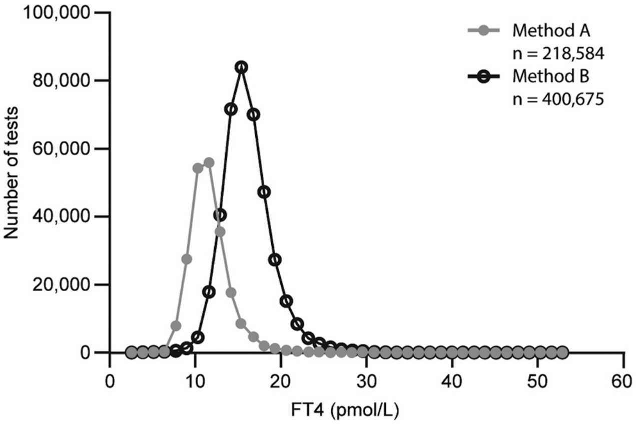

Figure 1: Distribution of FT4 results from routine tests across two commercial immunoassays (A and B) in 2017 (unpublished data provided by José Gilberto Henriques Vieira).
Internal quality control procedures were used to derive precision estimates for each assay (produced by EP Evaluator software, Data Innovations LLC). Besides internal QC, methods were submitted to external QC programs, the PELM: Programs of External Quality Control from the Brazilian Society of Pathology, and College of American Pathologists – CAP Surveys. Within-run and between-day coefficients of variation (respectively) for the method A assay were 6 and 3.9% at a mean value of 6.9 pmol/L, and 4.7 and 1.4% at a mean FT4 value of 28.5 pmol/L; corresponding estimates for the method B assay were 2.4 and 1.4% at a mean FT4 value of 11.7 pmol/L, and 1.6 and 0.9% at a mean FT4 value of 30.9 pmol/L. Normal ranges (5th–95th percentile values of cumulative data) were 7.7–16.6 pmol/L (method A) and 11.5–21.8 pmol/L (method B). FT4, free thyroxine; QC, quality control.

Here we review the path to standardization of FT4 assays, including a summary of the achievements in the process so far and our collective expert opinion on the challenges ahead.

Why standardize?

The core aim of immunoassay standardization is to establish metrological traceability and ensure that analytical results are comparable across assays, laboratories, and time. The standardization process requires a reference measurement system for use as calibration hierarchy; then, the calibrated assays used across different laboratories will provide measurements that are traceable to the top of the hierarchy within stated uncertainty constraints (Figure 2) [21].

The current lack of standardization of FT4 immunoassays leads to several challenges in the reporting and interpretation of patient test results. Patients may visit multiple health facilities that use different laboratories with varying testing methods for FT4, and clinicians from the separate facilities may discuss results without realizing that different methods have been used (as modifications to assay methodology are often not reported). Furthermore, over time, new measurement methods may be introduced, and laboratories may change the method used; for instance, when new analytical equipment is implemented. Global standardization of FT4 immunoassays would therefore result in a number of benefits, including the traceability of results back to a common standard, improved comparability of results between different platforms and vendors, and increased confidence in the use of common reference limits and/or decision limits for FT4 laboratory data in clinical decision-making.

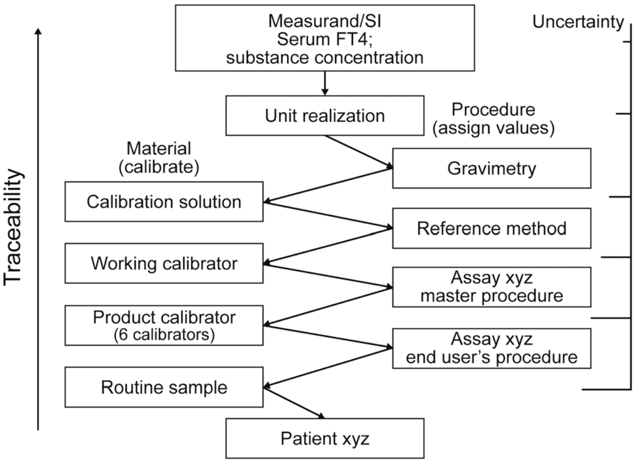

Figure 2: FT4 reference measurement system [21].
The C-STFT proposed an international "conventional" RMP for FT4 based on ED combined with direct determination of the thyroxine concentration in the dialyzate with a trueness-based RMP utilizing ID-LC/tandem MS. The proposed FT4 measurement system provides full metrological traceability from any given patient (patient *xyz*) and any given assay (assay *xyz*) back to the original reference materials. C-STFT, Committee for Standardization of Thyroid Function Tests; ED, equilibrium dialysis; FT4, free thyroxine; ID, isotope dilution; LC, liquid chromatography; MS, mass spectrometry; RMP, reference measurement procedure; SI, international system of units. Reproduced from the IFCC C-STFT article "Standardization of FT4 and FT3 measurements". https://ifcc-cstft.org/standardization-of-FT4-and-FT3-measurements [21]. Copyright 2019, with permission from the IFCC.

Standardization and reference limits

Reference intervals for FT4 are population-specific and can therefore vary considerably, even between laboratories using the same assay. One study reported a difference of 14% in the lower reference limit and 6% in the upper reference limit between an institution's own FT4 method-specific intervals and those quoted by the assay manufacturer [22]. Differences in reference intervals may be due to variations in reference population definitions (e.g., differences in inclusion/exclusion criteria based on laboratory parameters) or due to intrinsic differences between populations (e.g., ethnicity, diet, genetics). Some FT4 reference interval studies have only included FT4 values from healthy subjects with TSH concentrations within the reference intervals for TSH [23] thereby potentially introducing population bias. Moreover, manufacturers may use different percentile cut-off values to represent their lower and upper limits. Currently, the percentiles used for the reference intervals differ between covering the central 95% distribution (2.5th–97.5th percentile) to the central 99% distribution (0.5th–99.5th percentile) (Table 1) [24]. The former is recommended by the Clinical and Laboratory Standards Institute [25]; using a broader reference interval might result in inadequate detection of subclinical cases. Overall, although the clinical utility of different FT4 immunoassays can be addressed by using well-defined assay-specific reference intervals and/or clinical decision limits, this does not provide comparability of results between laboratories that use different testing platforms. In addition, using reference ranges and clinical decision limits proposed in literature without taking into account differences between specific assays may reduce confidence in clinical decision making.

The use of different reference populations, inclusion/exclusion criteria, and percentiles as cut-off values for reference intervals may also be addressed during the standardization process, leading to the establishment of universal, or at least population-specific, reference limits. These limits would be determined using results from target populations [26], although the definition of standard inclusion and exclusion criteria for the purposes of thyroid hormone assays must also be established.

It should be noted, however, that alterations in binding protein levels seen in different physiologic variables have varying impact on different manufacturer FT4 immunoassays. It may still require assay-specific interpretation or alternative approaches post-standardization in some physiological conditions such as pregnancy, drug interactions, or presence of binding protein inhibitors in non-thyroidal illness. Thus, standardization efforts should include diverse patient populations.

FT4 testing in pregnancy

FT4 testing in pregnancy is challenging [27]. Dynamic changes in thyroid function during normal pregnancy due to rising human chorionic gonadotropin levels, as well as a number of other pregnancy-related physiologic changes, such as alterations in albumin and thyroxine-binding globulin concentrations, can all affect FT4 immunoassays [27–30]. Serum FT4 concentrations are known to vary during pregnancy (Figure 3) [31], and several studies have indicated that non-pregnancy reference intervals of serum FT4 are not applicable for diagnosing thyroid diseases during pregnancy [32]. A longitudinal study of 130 healthy pregnant women reported that 36% of FT4 values recorded during the second trimester fell below the non-pregnancy lower normal limit, and this proportion rose to 41% in the third trimester [33]. It is generally accepted that non-pregnancy reference intervals can be used up to and including week 6 of gestation only [34].

Current FT4 immunoassays vary in their sensitivity to alterations in binding proteins that occur during pregnancy. A study of commercial immunoassays in pregnant and age-matched non-pregnant females showed that mean FT4 concentrations (measured by ED and MS) were 8.8% lower in pregnant females than non-pregnant controls in the late first trimester, and 29.1% lower in the second and third trimesters [28]. These findings were then compared with results from three commercial immunoassays: two of the three commercial assays produced similar results to the ED-MS analysis, but one showed no decrease in FT4 in the late first trimester and a less pronounced decrease in the second and third trimesters (15 and 14.4%, respectively). All the commercial assays were found to be affected by alterations in thyroxine-binding protein levels during pregnancy [28]. The 2017 American Thyroid Association guidelines for the management of thyroid disease during pregnancy acknowledge that the accuracy of FT4 immunoassays are dependent on trimester and assay procedure, and recommend that trimester- and assay-specific reference intervals should be applied when measuring FT4 in pregnant women [35]. While standardization will minimize differences in FT4 results across assays, inaccuracy of FT4 results due to different susceptibility of FT4 immunoassays to inaccuracies due to pregnancy-related alterations in binding proteins will not be addressed by standardization, so it is likely that challenges will remain in this population post-standardization.

The expert panel highlights that the pregnant population requires special consideration post FT4 assay standardization. Re-establishment of trimester- and method-specific reference intervals is crucial for the successful implementation of standardized FT4 testing. Furthermore, clinicians must be able to confidently interpret the results of gestational FT4 testing obtained with the new standardized method.

Table 1: Reference intervals for different manufacturer FT4 assays [24].

IVD manufacturer; platform/ immunoassay	Reference interval, pmol/L	Percentile
Siemens healthineers (Tarrytown, NY, USA); Advia Centaur XP	11.5–22.7	NR
Abbott Diagnostics (Abbott Park, IL, USA); ARCHITECT i2000	9.0–19.1	99%
Ortho-Clinical Diagnostics (Buckinghamshire, UK); Vitros ECi	10.0–28.2	98%
bioMerieux SA (Marcy-l'Etoile, France); Vidas	10.6–19.4	95%
Beckman Coulter Inc. (Brea, CA, USA); Access 2	7.9–14.4	95%
DiaSorin S.p.A (Saluggia, Italy); Liaison® Analyzer	10.3–21.9	95%
Sichuan Maccura Biotechnology Co., Ltd. (Chengdu, China); IS1200	12.2–21.2	95%
Roche Diagnostics International Ltd (Rotkreuz, Switzerland); Elecsys® cobas e 601	12.0–22.0	95%
Tosoh Corporation (Tokyo, Japan); AIA-2000	10.6–21.0	95%
Snibe Co., Ltd., (Shenzhen, China); Maglumi 2000	11.5–22.1	95%
Fujirebio Inc. (Tokyo, Japan); Lumipulse G1200	9.7–19.8	95%
LSI Medience Corporation (Tokyo, Japan); STACIA	12.5–26.5	NR
Sysmex Corporation (Kobe, Japan); HISCL-5000	9.9–20.5	NR

FT4, free thyroxine; IVD, *in vitro* diagnostic; NR, not reported. Modified from De Grande LAC, et al. Standardization of free thyroxine measurements allows the adoption of a more uniform reference interval. Clin Chem 2017;63:1642–52 [24]. http://www.clinchem.org/. Copyright 2019, with permission.

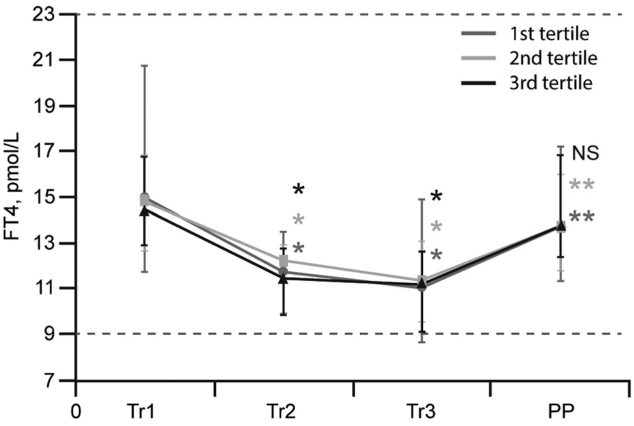

Figure 3: Serum FT4 concentrations measured by immunoassay during each trimester of pregnancy and postpartum (n=60) by TSH level in trimester 1 [31].
Error bars represent the 5th and 95th percentiles per trimester. Dashed lines represent non-pregnant lower and upper reference limits. All analyses were performed on a cobase 601 analyzer (Roche Diagnostics). *p<0.0001 vs. Tr1; **p<0.01 vs. Tr1; FT4, free thyroxine; NS, not significant; PP, postpartum; Tr, trimester; TSH, thyroid stimulating hormone. Reproduced from Joosen AM, et al. TSH and FT4 during pregnancy: an observational study and a review of the literature. Clinical Chemistry and Laboratory medicine 2016;54:1239–46 [31]. Copyright 2019, with permission.

Pediatric FT4 testing

Establishing reference intervals in pediatric subjects is particularly challenging due to the continuous physiologic changes that occur throughout childhood [36, 37]. Reference interval studies would need to include multiple age- and sex-specific partitions, requiring many samples to develop reliable estimates [36]. Most available pediatric reference intervals have applied statistical measures to hospital populations to provide estimates, but the establishment of accurate reference intervals obtained from a healthy pediatric population is crucial for the correct clinical interpretation of laboratory results. FT4 concentrations are higher at birth and decrease throughout the first year of life [36]. The most remarkable dynamic change in FT4 levels occurs during the first week of life, with concentrations rising immediately after birth and peaking at around 24 h (Figure 4) [38]. This means that the upper reference limit of FT4 may be substantially higher in neonates than in adults. Given the variation in FT4 concentrations with age, the expert panel recommends that pediatric reference intervals should be established. Global standardization of FT4 assays will allow collaborative efforts toward establishing these reference intervals.

FT4 testing in patients taking levothyroxine

In patients prescribed levothyroxine to treat hypothyroidism, the adequacy of thyroid hormone replacement is monitored by measuring TSH and/or FT4 levels. The treatment goal for primary hypothyroidism is to achieve TSH concentrations within the reference interval. However, the concentrations of FT4 that correspond to normal TSH values in this patient population have been shown to be higher than those in euthyroid subjects [39]. A study of samples from patients taking levothyroxine showed a clear shift to the right in the FT4 frequency distribution curve compared with control patients with no known thyroid disease (Figure 5), and reported that 10.3% of the levothyroxine-treated patients had FT4 concentrations above the manufacturer's upper reference limit, alongside a normal TSH measurement [40]. As FT4 standardization is expected to increase numeric FT4 results, measurements for patients taking levothyroxine should consider the expected higher FT4 levels in these patients. Increasing the upper limit for this population would reduce inappropriate flagging of FT4 results and prevent unnecessary levothyroxine dose adjustment.

The expert panel recommends that specific intervals for patients on levothyroxine therapy should be considered during the FT4 standardization process. However, it is important to note that patient-related factors such as variations in thyroid hormone receptor sensitivity, age, and

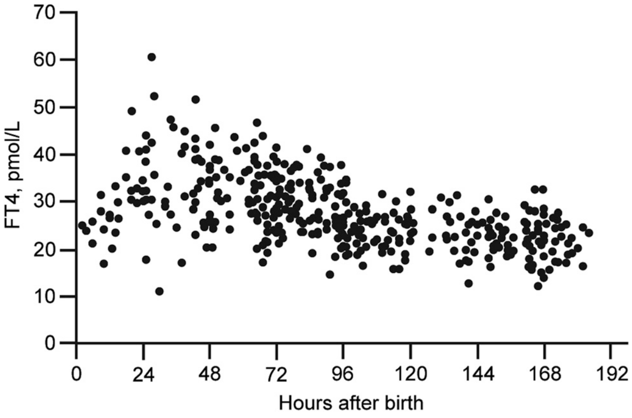

Figure 4: Distribution of FT4 concentrations in the first 7 days of life in neonates without known thyroid disease with one episode of thyroid function testing; FT4 concentrations rise immediately after birth and peak at around 24 h [38].
FT4, free thyroxine. Reproduced from Jayasuriya MS, et al. Reference intervals for neonatal thyroid function tests in the first 7 days of life. Journal of Pediatric Endocrinology & Metabolism 2018;31:1113–6 [38]. Copyright 2019, with permission.

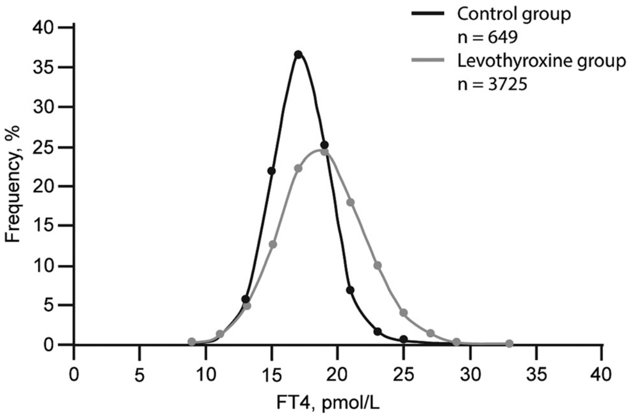

Figure 5: Distribution of FT4 in people taking levothyroxine (n=3725) and in normal controls without known thyroid disease (n=649).
There is a clear shift to the right in the FT4 frequency distribution curve in the levothyroxine group compared with control patients with no known thyroid disease [40]. FT4, free thyroxine. eproduced from Lu ZX, et al. Should there be separate free thyroxine reference limits for thyroxine-treated patients? The Clinical Biochemist Reviews 2016;37:S40 [40]. Copyright 2019, with permission.

gender add further complexity to the challenge of establishing limits in these patients. In practice, most patients in this setting would be monitored via their TSH levels and only those with central hypothyroidism would require FT4 monitoring.

The road to standardization: what has been achieved so far?

In collaboration with the *in vitro* diagnostic (IVD) industry, the IFCC C-STFT initiated an evaluation of the quality and comparability of commercial FT4 assays [41]. These method comparison studies demonstrated that the assays were of good quality but reported differences between assay results [42]. However, the comparison studies mostly used samples from healthy donors and few samples from patients with thyroid disease. Therefore, the C-STFT undertook a further method comparison study using a wide variety of samples from patients with thyroid disease to explore the potential impact of standardization across the clinically relevant FT4 concentration range [19]. The authors reported that all of the 13 tested FT4 assays were negatively biased in the mid-to-high concentration range, with a maximum inter-assay discrepancy of around 30%. In the low concentration range, the maximum inter-assay discrepancy was approximately 90%. Recalibration had a substantial impact on the tested assays; inter-assay differences were eliminated, and the remaining data dispersion was almost entirely due to within-assay error (Supplemental Figure 1) [19].

As standardization relies on a robust RMP, it is critical to optimize the RMP and control for known analytical challenges. The C-STFT has developed an international conventional RMP for the measurement of FT4 concentrations in serum using pmol/L units at physiologic pH 7.40 and temperature 37 °C, to ensure that the measurement accurately reflects the concentration of FT4. For free thyroid hormones, the measurement procedure must include a physical step for separation of the free hormone from bound fraction (e.g., by ED). Due to this separation step, it is not possible to unequivocally show that the thyroxine concentration in dialyzate is identical to the true free hormone concentration in the original sample, even if ED is performed under ideal physiologic conditions. The C-STFT therefore proposed a conventional RMP for FT4 based on ED ID-LC/tandem MS. The convention refers to the ED part of the RMP, which must strictly adhere to a predefined procedure [21, 43].

The proposed RMP for FT4 testing is being validated and optimized by a network of reference laboratories in Belgium, The Netherlands, Japan, and the USA. A study involving 13 different commercial FT4 assays recalibrated against the RMP showed that recalibration significantly reduced FT4 immunoassay bias, and that the reference interval determined by the RMP was suitable for common use within a margin of 12.5% (Supplemental Figure 2) [23]. This outcome represented substantial progress in the standardization of FT4 measurements.

The road to standardization: what next?

The work of the C-STFT project on FT4 assay recalibration has advanced the field significantly, but substantial work is still required before global standardization can be achieved. The current RMP for FT4 testing is technically demanding and longitudinal consistency is challenging; work is ongoing across the reference laboratories to address the technical issues and optimize the procedure. The IFCC C-RIDL have identified various challenges for harmonizing reference intervals and have sequentially determined several strategies to overcome them [44]. Ongoing projects include: a study to compare alternative approaches for the determination of reference intervals; a website to provide the reference intervals obtained from a global study performed by C-RIDL for practice of Evidence Based Laboratory Medicine; and a publication on the distinction of Reference Intervals and Clinical Decision Limits [44]. Additionally, there are various national initiatives ongoing in countries such as Canada, Ireland, the United States of America, and South Korea, to establish better reference intervals [45–48]. Reference value studies will have to be performed in several different populations, including adults, children, neonates, pregnant women, and people taking levothyroxine. Age-adjusted reference intervals may also be necessary. Studies to provide comparisons of current standardized IVD company methods with the C-STFT reference method will also be required, and the recalibration equation for each manufacturer will have to be certified and monitored for stability over time. Crucially, the clinical decision limits used in practice guidelines will have to be re-evaluated.

The regulatory requirements associated with the recalibration of assays will also have to be considered before standardization can be implemented. The C-STFT has contacted major regulatory agencies to establish what they will require from IVD manufacturers who recalibrate their assays. New published guidelines for FT4 testing will be required following standardization, and country-specific guidance will also have to be updated. The expert panel acknowledges that there is a risk that the changes in numeric measurement values and reference intervals after standardization could result in misinterpretation of laboratory data. Therefore, a major challenge in the process of standardization will be to educate and prepare laboratories, clinicians, and patients for the new methods in FT4 testing and the changes in FT4 results. As a first step in the process of education, the C-STFT has contacted national laboratory societies, general practitioners, endocrinologists, thyroidologists, laboratory specialists, nurses, and patient organizations to create awareness of the standardization process, summarize the achieved milestones, and to seek their input [20]. The responding stakeholders supported the need for standardization but highlighted the potential risks to patient safety and clinical outcomes arising from major changes in numeric values and reference intervals.

The overall recommendation of the expert panel is that education of FT4 standardization should be at three levels, to be overseen by an international working party led by the IFCC: (i) guidelines and expert recommendation published in journals; (ii) congress communication; (iii) laboratory communications in the local community.

IVD manufacturers have a crucial role in working together to implement the change in a coordinated manner. They should also take responsibility for providing literature to laboratories to explain the standardization process in terms of why it is necessary, when the changes will come into effect, and how the transition will be handled. It would also be desirable for IVD companies to facilitate user group meetings to help laboratory professionals understand the changes and enable them to communicate the information to healthcare professionals. For successful standardization, the process will have to be discussed, shared, and promoted via national and international societies for laboratory medicine and clinical endocrinology. Journal articles and meetings organized by scientific societies will be crucial to explain the changes and illustrate the differences in FT4 numeric values before and after standardization. Inter-society and interdisciplinary meetings will ensure that the information is disseminated effectively across all stakeholders. At a local level, case studies could be presented and discussed to illustrate the changes. It is important that these educational programs also include information about what standardization will not achieve (e.g. binding protein effects or interferences due to factors present in samples from individual patients, including biotin, will not be solved by standardization). Concise pocket guides or webpages containing the most important practical information (e.g., old and new reference intervals) may be useful for stakeholders while they are adjusting to the new methods. Patient education about FT4 standardization should primarily be provided by clinicians who can explain what it means for the patient in terms of numeric changes in FT4 concentration results. A simple explanatory booklet should also be considered for distribution via healthcare professionals and patient associations. A key strategy would be the development of an educational FT4 standardization website that caters for all the different stakeholders (i.e., manufacturers, laboratories, clinicians and other healthcare professionals, researchers, and patients).

How should the transition phase be handled?

The expert panel recognizes that the changes associated with FT4 standardization may pose a particular risk to patient care during the transition phase. The standardized FT4 values will be significantly higher than current values, meaning that the numeric FT4 value for the upper limit of normal after standardization will be above the current critical limits for clinical action. To mitigate problems during the transition phase, clinicians need to be fully informed of the changes, and the expert panel recommends that laboratory experts should explain the changes to clinicians.

During the transition phase, FT4 results could be expressed in both old and new values with the old and new reference intervals included on the patient reports. Decisions on how best to report results during the transition phase will likely be a matter for discussion between laboratories and clinicians; however, in either case, conversion factors should be provided to facilitate comparisons between old and new test results.

Has assay standardization been successfully achieved before?

When contemplating the standardization process for FT4 testing, it is prudent to review previous examples of assay standardization to mitigate potential pitfalls and plan for the challenges ahead. Assays for glycated hemoglobin (measured as HbA1c) became widely used for diabetes monitoring in the mid-1990s, but there were significant differences in test results between laboratories due to the diverse range of methods being used and the lack of a primary reference material [44].

National HbA1c standardization programs were established in the USA, Japan, and Sweden, but the lack of internationally recognized and accepted reference materials and procedures meant that accuracy and comparability of HbA1c measurements could not be guaranteed at a global level [49]. In 1994, the IFCC established a Working Group on HbA1c Standardization and developed two equivalent reference methods as well as primary reference materials [50]. The reference methods were accepted by the National Societies of Clinical Chemistry and published in 2002 [51]. In 2004, the IFCC recommended that all manufacturers of equipment used in HbA1c assays should calibrate their methods to the IFCC reference methods [52].

HbA1c standardization significantly changed the numeric results provided to clinicians, causing concern from specialists that reporting the standardized HbA1c values might lead to misinterpretation of the degree of glycemic control and cause confusion for clinicians and patients. It was feared that during the transition phase, there would be a worsening of glycemic control, resulting in adverse clinical outcomes for patients [50].

Almost 15 years after the acceptance of the IFCC reference methods, assay manufacturers have now calibrated their HbA1c assays to the highest international standard, although adoption of the new unit and reference intervals is still slow outside of Europe [52]. In the USA, the National Glycohemoglobin Standardization Program (NGSP) has led standardization efforts since 1996 [53]. The NGSP and IFCC approaches to the standardization of HbA1c results serve different, but complementary, purposes. The primary objective of IFCC standardization is to ensure that manufacturers are traceable to an accuracy base, but there is no limit on the degree of uncertainty allowed between a manufacturer's method and a reference method-assigned value. The NGSP also defines acceptable limits for method performance that are based on clinical need; for example, recommendations for diabetes care by clinical societies [53]. There is now an ongoing collaboration between the IFCC and NGSP to ensure that the relationship between the two networks remains consistent over time.

HbA1c results are still commonly being reported in different units (mmol/mol vs. percent), but standardization has reduced assay bias and HbA1c results are now more comparable between different laboratories [52]. As a result, a single common cut-off value of 48 mmol/mol (6.5%), regardless of the assay methodology is accepted as one of the diagnosis criteria for diabetes [52].

The HbA1c standardization project has faced substantial challenges in implementing the IFCC recommendations globally, despite effectively engaging with clinicians, biochemists, external quality assessment organizers, patient groups, and manufacturers to undertake a large-scale educational program [49]. Accordingly, a deadline for maintaining both old and new reference intervals should be defined in order to avoid differences and confusion. Several differences may help expedite FT4 standardization, compared with efforts applied to the HbA1c test. There is only one established standardization protocol, and standardization will not change the units used to express FT4 test results. Although a variety of different approaches are used, most commercially available FT4 assays are immunoassays, whereas a variety of very different technologies are used to measure HbA1c. On the other hand, it should be considered that the different approaches used for the measurement of free hormones (i.e., one step, two step, homogeneous phase, etc.) entail considerable differences in the sensitivity to different thyroid conditions for which the measurement is carried out. FT4 has less clinical significance and is probably less commonly requested than HbA1c. Consequently, the use of accuracy grading in proficiency testing, which was successful in aligning HbA1c assays [54], as well as the adoption of educational programs, will probably be more effective in achieving standardization.

Summary

The expert panel agreed that standardized immunoassays are required to address modern clinical and public health needs, and to increase healthcare professionals' confidence in safely using laboratory data for clinical decision-making. Standardization of FT4 testing will facilitate accurate interpretation of laboratory thyroid function data, thus ensuring optimal patient care.

The panel also acknowledged that standardized FT4 testing will translate into changed clinical decisions in a relatively small minority of cases, meaning that some stakeholders may question whether the effort and resources required for standardization outweigh the clinical benefit. It is essential that all stakeholders are engaged and aligned for standardization to be successfully achieved.

Conclusions

Significant progress has already been made in the standardization of procedures for FT4 testing, but technical and implementational challenges remain, including establishment of clinical decision limits in different patient populations and education of all stakeholders. The experiences of previous standardization programs give a valuable insight into the potential problems that may arise and allow us to plan strategies to overcome them. Without strong involvement from clinical societies or the adoption of clinical guidelines and standards in the endocrinology community, the education and acceptance of standardized FT4 values will not work.

Acknowledgments: The authors wish to thank Michael Rottmann, PhD, and James Faix, MD, for their contribution to the development of the manuscript. Third-party medical writing assistance, under the direction of the authors, was provided by Angela Corstorphine and Sian-Marie Lucas of Kstorfin Medical Communications Ltd, and by Lucy Carrier of Gardiner-Caldwell Communications. COBAS, COBAS E, and ELECSYS are trademarks of Roche. All other product names and trademarks are the property of their respective owners.

Research funding: This study was funded by Roche Diagnostics International Ltd, Rotkreuz, Switzerland.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved its submission.

Competing interests: JGHV owns stock in Grupo Fleury. DK is an employee of Roche Diagnostics International Ltd. All authors participated in an advisory board for Roche Diagnostics; Zhong Lu did not accept any honorarium for the advisory board. The sponsor was involved in study design, the collection and interpretation of the data, and writing of the manuscript.

References

1. Toubert ME, Chevret S, Cassinat B, Schlageter MH, Beressi JP, Rain JD. From guidelines to hospital practice: reducing inappropriate ordering of thyroid hormone and antibody tests. Eur J Endocrinol 2000;142:605–10.
2. Schneider C, Feller M, Bauer DC, Collet TH, da Costa BR, Auer R, et al. Initial evaluation of thyroid dysfunction – are simultaneous TSH and fT4 tests necessary? PloS One 2018;13:e0196631.
3. Viera AJ. Thyroid function testing in outpatients: are both sensitive thyrotropin (sTSH) and free thyroxine (FT4) necessary? Fam Med 2003;35:408–10.
4. Biondi B, Bartalena L, Cooper DS, Hegedus L, Laurberg P, Kahaly GJ. The 2015 European Thyroid Association guidelines on diagnosis and treatment of endogenous subclinical hyperthyroidism. Eur Thyroid J 2015;4:149–63.
5. Ross DS, Burch HB, Cooper DS, Greenlee MC, Laurberg P, Maia AL, et al. American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. Thyroid 2016;2016:1343–421.
6. Redford C, Vaidya B. Subclinical hypothyroidism: should we treat? Post Reprod Health 2017;23:55–62.
7. Persani L, Brabant G, Dattani M, Bonomi M, Feldt-Rasmussen U, Fliers E, et al. European Thyroid Association (ETA) guidelines on the diagnosis and management of central hypothyroidism. Eur Thyroid J 2018;2018:225–37.
8. LeFevre ML. Screening for thyroid dysfunction: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2015;162:641–50.
9. Garber JR, Cobin RH, Gharib H, Hennessey JV, Klein I, Mechanick JI, et al. Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American association of clinical endocrinologists and the American thyroid association. Endocr Pract 2012;18: 988–1028.
10. Beck-Peccoz P, Lania A, Beckers A, Chatterjee K, Wemeau JL. European Thyroid Association guidelines for the diagnosis and treatment of thyrotropin-secreting pituitary tumors. Eur Thyroid J 2013;2013:76–82.
11. Rivas AM, Lado-Abeal J. Thyroid hormone resistance and its management. In: Proc (Bayl Univ Med Cent). Waco, Texas, USA: Baylor University Medical Center; 2016, vol 29:209–11 pp.
12. Schussler GC. The thyroxine-binding proteins. Thyroid 2000;10: 141–9.
13. van Deventer HE, Soldin SJ. The expanding role of tandem mass spectrometry in optimizing diagnosis and treatment of thyroid disease. Adv Clin Chem 2013;61:127–52.
14. Revet I, Boesten LSM, Linthorst J, Yildiz E, Janssen JW, de Rijke YB, et al. Misleading FT4 measurement: assay-dependent antibody interference. Biochem Med 2016;26:436–43.
15. Gounden V, Jonklaas J, Soldin SJ. A pilot study: subclinical hypothyroidism and free thyroid hormone measurement by immunoassay and mass spectrometry. Clin Chim Acta 2014;430: 121–4.
16. van Deventer HE, Mendu DR, Remaley AT, Soldin SJ. Inverse log-linear relationship between thyroid-stimulating hormone and

free thyroxine measured by direct analog immunoassay and tandem mass spectrometry. Clin Chem 2011;57:122–7.

17. Pop VJ, Brouwers EP, Vader HL, Vulsma T, van Baar AL, de Vijlder JJ. Maternal hypothyroxinaemia during early pregnancy and subsequent child development: a 3-year follow-up study. Clin Endocrinol 2003;59:282–8.
18. Mimoto MS, Refetoff S. Clinical recognition and evaluation of patients with inherited serum thyroid hormone-binding protein mutations. J Endocrinol Invest 2020;43:31–41.
19. Thienpont LM, Van Uytfanghe K, Van Houcke S, Das B, Faix JD, MacKenzie F, et al. A progress report of the IFCC committee for standardization of thyroid function tests. Eur Thyroid J 2014;3: 109–16.
20. Thienpont LM, Faix JD, Beastall G. Standardization of FT4 and harmonization of TSH measurements – a request for input from endocrinologists and other physicians. Endocrine 2015;62: 855–6.
21. IFCC. Standardization of FT4 and FT3 measurements. Available from: https://ifcc-cstft.org/standardization-of-FT4-and-FT3-measurements [Accessed 28 Feb 2020].
22. Lee GR, Griffin A, Halton K, Fitzgibbon MC. Generating method-specific reference ranges – a harmonious outcome? Pract Lab Med 2017;9:1–11.
23. Yeap BB, Manning L, Chubb SA, Hankey GJ, Golledge J, Almeida OP, et al. Reference ranges for thyroid-stimulating hormone and free thyroxine in older men: results from the Health in Men Study. J Gerontol A Biol Sci Med Sci 2017;72:444–9.
24. De Grande LAC, Van Uytfanghe K, Reynders D, Das B, Faix JD, MacKenzie F, et al. Standardization of free thyroxine measurements allows the adoption of a more uniform reference interval. Clin Chem 2017;63:1642–52.
25. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory – CLSI document EP28-A3C, 3rd ed. Philadelphia (PA): Clinical and Laboratory Standards Institute; 2008.
26. Baloch Z, Carayon P, Conte-Devolx B, Demers LM, Feldt-Rasmussen U, Henry JF, et al. Laboratory medicine practice guidelines. Laboratory support for the diagnosis and monitoring of thyroid disease. Thyroid 2003;13:3–126.
27. Lee RH, Spencer CA, Mestman JH, Miller EA, Petrovic I, Braverman LE, et al. Free T4 immunoassays are flawed during pregnancy. Am J Obstet Gynecol 2009;200:260.e261–6.
28. Anckaert E, Poppe K, Van Uytfanghe K, Schiettecatte J, Foulon W, Thienpont LM. FT4 immunoassays may display a pattern during pregnancy similar to the equilibrium dialysis ID-LC/tandem MS candidate reference measurement procedure in spite of susceptibility towards binding protein alterations. Clin Chim Acta 2010;411:1348–53.
29. Glinoer D. The regulation of thyroid function in pregnancy: pathways of endocrine adaptation from physiology to pathology. Endocr Rev 1997;18:404–33.
30. Moleti M, Trimarchi F, Vermiglio F. Thyroid physiology in pregnancy. Endocr Pract 2014;20:589–96.
31. Joosen AM, van der Linden IJ, de Jong-Aarts N, Hermus MA, Ermens AA, de Groot MJ. TSH and fT4 during pregnancy: an observational study and a review of the literature. Clin Chem Lab Med 2016;54:1239–46.
32. Gao X, Li Y, Li J, Liu A, Sun W, Teng W, et al. Gestational TSH and FT4 reference intervals in Chinese women: a systematic review and meta-analysis. Front Endocrinol 2018;9:432.
33. Ekinci EI, Lu ZX, Sikaris K, Bittar I, Cheong KY, Lam Q, et al. Longitudinal assessment of thyroid function in pregnancy. Ann Clin Biochem 2013;50:595–602.
34. Laurberg P, Andersen SL, Hindersson P, Nohr EA, Olsen J. Dynamics and predictors of serum TSH and fT4 reference limits in early pregnancy: a study within the Danish National Birth Cohort. J Clin Endocrinol Metab 2016;101:2484–92.
35. Alexander EK, Pearce EN, Brent GA, Brown RS, Chen H, Dosiou C, et al. Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. Thyroid 2017;2017:315–89.
36. Bailey D, Colantonio D, Kyriakopoulou L, Cohen AH, Chan MK, Armbruster D, et al. Marked biological variance in endocrine and biochemical markers in childhood: establishment of pediatric reference intervals using healthy community children from the CALIPER cohort. Clin Chem 2013;59: 1393–405.
37. Kratzsch J, Schubert G, Pulzer F, Pfaeffle R, Koerner A, Dietz A, et al. Reference intervals for TSH and thyroid hormones are mainly affected by age, body mass index and number of blood leucocytes, but hardly by gender and thyroid autoantibodies during the first decades of life. Clin Biochem 2008;41:1091–8.
38. Jayasuriya MS, Choy KW, Chin LK, Doery J, Stewart A, Bergman P, et al. Reference intervals for neonatal thyroid function tests in the first 7 days of life. J Pediatr Endocrinol 2018;31:1113–6.
39. Galior K, Stan M, Baumann N. Higher FT4 results in levothyroxine-treated patients with normal TSH compared to patients without thyroid disease. J Endocr Soc 2019;3:MON-623.
40. Lu ZX, Sikaris KA, Yen T, Trambas C, Walsh J. Should there be separate free thyroxine reference limits for thyroxine-treated patients? Clin Biochem Rev 2016;37:S40.
41. IFCC. Standardization of thyroid function tests. Available from: http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-stft/ [Accessed 28 Oct 2020].
42. Thienpont LM, Van Uytfanghe K, Van Houcke S. Standardization activities in the field of thyroid function tests: a status report. Clin Chem Lab Med 2010;48:1577–83.
43. Van Houcke SK, Van Uytfanghe K, Shimizu E, Tani W, Umemoto M, Thienpont LM. IFCC international conventional reference procedure for the measurement of free thyroxine in serum: International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Working Group for Standardization of Thyroid Function Tests (WG-STFT)(1). Clin Chem Lab Med 2011;49: 1275–81.
44. IFCC. Reference intervals and decision limits (C-RIDL). Available from: https://www.ifcc.org/ifcc-scientific-division/sd-committees/c-ridl/ [Accessed 28 Oct 2020].
45. AACC. What is the current state of reference intervals in Canada? A need for harmonization. Available from: https://www.aacc.org/science-and-research/scientific-shorts/2017/current-state-of-reference-intervals-in-canada [Accessed 28 Oct 2020].
46. McCafferty R, McHugh J, Regan I, Mannion A, Boran G. National Clinical Programme For Pathology Reference Interval Harmonisation Project Group: 2nd draft survey report on reference intervals for the full blood count in the Republic of Ireland; 2017. Available from: http://rcpi-live-cdn.s3.amazonaws.com/wp-content/uploads/2017/09/Reference-Intervals-for-the-FBC-in-ROI-for-NCPP-2017.pdf [Accessed 28 Oct 2020].

47. Katayev A, Balciza C, Seccombe D. Establishing reference intervals for clinical laboratory test results: is there a better way? Am J Clin Pathol 2010;133:180–6.
48. Park SY, Kim HI, Oh HK, Kim TH, Jang HW, Chung JH, et al. Age- and gender-specific reference intervals of TSH and free T4 in an iodine-replete area: data from Korean National Health and Nutrition Examination Survey IV (2013–2015). PloS One 2018;13:e0190738.
49. John WG, Mosca A, Weykamp C, Goodall I. HbA1c standardisation: history, science and politics. Clin Biochem Rev 2007;28:163–8.
50. Lai LC. Global standardisation of HbA1c. Malays J Pathol 2008;30: 67–71.
51. Jeppsson JO, Kobold U, Barr J, Finke A, Hoelzel W, Hoshino T, et al. Approved IFCC reference method for the measurement of HbA1c in human blood. Clin Chem Lab Med 2002;40:78–89.
52. Penttilä I, Penttilä K, Holm P, Laitinen H, Ranta P, Törrönen J, et al. Methods, units and quality requirements for the analysis of haemoglobin A1c in diabetes mellitus. World J Methodol 2016;6: 133–42.
53. Little RR, Rohlfing C, Sacks DB. The National Glycohemoglobin Standardization Program: over 20 years of improving hemoglobin A1c measurement. Clin Chem 2019;65:839–48.
54. Little RR, Rohlfing CL, Sacks DB. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. Clin Chem 2011;57:205–14.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2020-1696).